Reading As Humans: Unified Negation Detection via Cascade Linear Attention Network with Model-Agnostic Meta-Learning

Zhong Qian✉[1] **Peifeng Li**[1] **Qiaoming Zhu**[1] **Guodong Zhou**[1]

[1]School of Computer Science and Technology, Soochow University, Suzhou 215006, China

Abstract

Negation Detection (ND) aims to extract negative statements with non-factual semantics, including negative scopes, xscopes, focuses, where (x)scopes are continuous spans governed by negative cues, and focuses are the most principal negated texts in scopes. Previous studies are confined to task-specific models, and failed to develop a unified solution for (x)scopes and focuses based on the correlations of their definitions, nor implemented the mutual promotion for each task. Moreover, existing work did not design fine-grained encoders, resulting in reliance on external features, such as syntactic parse trees. To address these issues, we construct a unified negation detection paradigm with Model-Agnostic Meta-Learning (MAML) to facilitate those ND sub-tasks. Besides, inspired by reading mechanism of humans, we synthesize Cascade Linear Attention Network (CLAN) by heterogeneous linear attentions with cascade coefficients as fine-grained encoder. Experimental results on several corpora demonstrate that CLAN can outperform the state-of-the-arts.

1 INTRODUCTION

Negation Detection (ND) aims to extract those negated statements that convey counterfactual semantics. The elements of negative statements mainly include the following types as illustrated by Figure 1. **Negative Cue** is a token or phrase that signals negative semantics, e.g., "*not*" in S1/S1.1/S1.2, "*no*" in S1.3. Generally speaking, a negative cue is the most significant lexical clue that can govern the following spans: 1) **Negative Scope/XScope** are continuous spans affected by a corresponding cue. The principal difference between them is that a scope usually incorporates a cue in it. But an xscope indicates an implicit and context-derived scope involving a cue outside it. For example, In S1/S1.1/S1.2, the negative scope is "*the doctor does not allow him to be discharged from the hospital*", which is governed by the cue "*not*". In S1.3, however, no scope can be found around the negative cue "*no*", but we can deduce that the negated text in the last clause is "*he would be discharged from the hospital this month*" as a discrete scope, i.e., xscope; 2) **Negative Focus** is the most prominent and precise negated texts in the scope. In S1, we have known that the cue "*not*" negates the truth of "*the doctor allows him to be discharged from the hospital*". However, we cannot decide the focus of "*not*" due to the lack of positive semantics in contexts of special scenarios. In terms of S1.1, according to the affirmative statement "*to take a walk in the inpatient building*" that is a "to do" infinitive phrase, the focus is "*to be discharged from the hospital*". While for S1.2, based on the positive text "*next month*" that is the time when "*Mr. Simpson leaves/is discharged from the hospital*", the focus should also be a temporal adverbial, i.e., "*this month*".

S1: *Due to Mr. Simpson's current health status, the doctor does not allow him to be discharged from the hospital.*

S1.1: *Due to Mr. Simpson's health status, the doctor does not allow him to be discharged from the hospital , but permits him to take a walk in the inpatient building .*

S1.2: *Due to Mr. Simpson's health status, the doctor does not allow him to be discharged from the hospital this month , and insists that he should receive another course of treatment before leaving the hospital next month .*

S1.3: *Although Mr. Simpson hopes he would be discharged from the hospital this month, the doctor still says no.*

Figure 1: Example sentences containing negation. *Sentences* are in *italics*, *negative cues* are *underlined*, *scopes* and *xscopes* are marked in *red* and *green*, *focuses* and corresponding *positive/affirmative texts* are highlighted in *red* and *blue* background, respectively.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5556–5572.

Therefore, ND is comprised of two primary tasks, i.e., **Negative Scope Resolution** (NSR) and **Negative Focus Detection** (NFD), and an auxiliary task **Negative XScope Resolution** (NXSR) due to the minority of xscopes in datasets. The vast majority of work so far concentrates on NSR and NFD. 1) Neural networks have become the mainstream solutions for NSR, e.g., LSTM with lexical features [Fancellu et al., 2016], CNN [Qian et al., 2016] and recursive network [Fei et al., 2020] with syntactic features, BERT with data augmentation [Qian et al., 2023, 2024]; 2) NFD models have also switched to network-based methods recently, such as word-/topic-level contextual attention network [Shen et al., 2019], hierarchical LSTM integrated with different levels of features [Hossain et al., 2020]; 3) Although most studies consider extracting scopes and focuses separately, several recent research tries to address the issues of cross-domain adaptation by merging several datasets [Yoshida et al., 2024], applying data augmentation [Qian et al., 2023, 2024], bringing in cross-task features [Hossain et al., 2020].

However, these latest work still have not solved these problems: 1) *Lack of Unified Model*. Most previous work only developed task-specific models for each sub-task of ND, i.e., detecting (x)scopes and focuses separately, and failed to construct a unified model to detect them concurrently. As mentioned above, the similar definitions of (x)scopes and focuses can be conducive to the detection of them mutually, and this clue is neglected by existing work; 2) *Coarse-grained Encoding*. Previous studies employed simple networks (e.g., CNN, LSTM, residual networks), leading to the dependence on external information, such as syntactic trees. Thus, they failed to design fine-grained encoders to capture high-level semantics of negative statements.

This paper is devoted to solve the above disadvantages, and proposes Cascade Linear Attention Network (CLAN) as the solution. In principle, our major contributions can be summarized as: 1) This is the first work that develops a unified framework for Negation Detection (ND) to extract negative (x)scopes and focuses simultaneously; 2) Our model is integrated with Model-Agnostic Meta-Learning (MAML) technique that can implement mutual promotion for subtasks of ND; 3) Our model adopts the encoders based on human-like reading mechanism, i.e., utilizes heterogeneous Linear Attentions (LA) with cascade attention coefficients, which can learn fine-grained and higher-level information of negation; 4) Experimental results show that our model is superior to state-of-the-arts on several datasets.

2 RELATED WORK

Negative Scope Resolution. Currently, most work is devoted to neural networks, and preliminary research employed simple variants, including CNN [Qian et al., 2016], LSTM [Fancellu et al., 2016, 2017, Zhang et al., 2021], CRF [Li and Lu, 2018], recursive networks [Fei et al., 2020, McKenna and Steedman, 2020], hybrid models [Lazib et al., 2020]. To name a few, Fancellu et al. [2016] utilized an LSTM network based on word & PoS embeddings. Qian et al. [2016] proposed a CNN working on paths between cues and tokens in syntactic trees, and Lazib et al. [2020] combined LSTM & CNN to capture knowledge from syntactic paths. Li and Lu [2018] designed a latent variable CRF for capturing long distance dependencies and labelling tokens. Fei et al. [2020] employed a recursive network sequence labeling model to learn dependency features.

Recent research utilized BERT family. Some work [Sergeeva et al., 2019, Khandelwal and Sawant, 2020, Zhao and Bethard, 2020] directly applied BERT. Other studies tried to introduce task-specific modules or more mechanisms. Hartmann and Søgaard [2021] exploited a sequence labeling model for multilingual texts with zero-shot cross-lingual transfer learning. Based on the positions of span boundaries, Wu and Sun [2023] engaged boundary shift loss to refine predicted boundaries. Qian et al. [2023, 2024] developed data augmentation frameworks by token-level replacement. To tackle different annotations, Yoshida et al. [2024] converted scopes of BioScope [Vincze et al., 2008] & SFU [Konstantinova et al., 2012] to those of Sherlock [Morante and Daelemans, 2012, Morante and Blanco, 2012], and merged them into a unified dataset with pre-trained BERT/RoBERTa as baselines. Christen et al. [2024] released a new corpus of annotated legal documents, and improved NSR in both zero-shot and multilingual settings by cross-lingual pre-trained language models.

The above work has already tried to construct solutions for cross-lingual/-domain NSR, but failed to build a higher-level unified model with better generalization ability that can bring in other negative elements (e.g., negative focuses) as augmentation and associate N(X)SR & NFD together.

Negative Focus Detection has experienced three stages: 1) *Rule-Based Systems*. Rosenberg and Bergler [2012] detected negative cues, scopes, and focuses with a heuristics system, whose rules are designed based on cues, part-of-speeches, and passive voice; 2) *Classic Machine Learning Methods*. Blanco and Moldovan [2012] exploited a supervised model relying on features of lexical information, constituent parse trees, semantic roles. Zou et al. [2014, 2015] developed topic-level graph models to enrich intra-sentence features with inter-sentence information from lexical and topic perspectives; 3) *Neural Networks*. Shen et al. [2019] devised LSTM-CRF networks integrated with contextual attentions on words and topics to encode sequential and long-range context dependency. Hossain et al. [2020] proposed multi-layer LSTM+CRF network enhanced by contextual sentences, semantic roles, negative scopes. Some of the above studies have considered negative scopes as external indicators. However, there is still no unified model for negation detection that can implement promotions for extracting (x)scopes and focuses simultaneously.

3 METHOD: CASCADE LINEAR ATTENTION NETWORK

Overall architecture of Cascade Linear Attention Network is presented in Figure 2, which is comprised of Input Layer (§ 3.2), Cascade Linear Attention Layer (§ 3.3), Output Layer (§ 3.4). To describe the structure more clearly, we firstly introduce various Cascade Linear Attentions (§ 3.1).

3.1 CASCADE LINEAR ATTENTIONS

To simulate the rapid reasoning capability of human brains and improve computational efficiency, we employ **Linear Attentions** (**LA**) as the backbone of the encoder: $\boldsymbol{Q}' = \mathrm{LA}(\boldsymbol{Q}, \boldsymbol{K})$, where $\boldsymbol{V} = \boldsymbol{K}$ is omitted. In multi-head version, $\boldsymbol{Q}_l$/$\boldsymbol{K}_l$ in the l-th LA are projected into $\{\boldsymbol{Q}_{m,l}\}_{m=0}^{M}$/$\{\boldsymbol{K}_{m,l}\}_{m=0}^{M}$. For each head, elements of attention weights are computed as $\boldsymbol{A}_{m,l}^{KQ}(i,j) = \alpha + \boldsymbol{B}_{m,l}^{KQ}(i,j)$, where α is bias, and $\boldsymbol{B}_{m,l}^{KQ}$ is attention coefficient approximated as the candidate memory $\tilde{\boldsymbol{B}}_{m,l}^{KQ}$ with linear calculation β:

$$\begin{aligned}\tilde{\boldsymbol{B}}_{m,l}^{KQ}(i,j) &= \beta(\boldsymbol{k}_i^{m,l}, \boldsymbol{q}_j^{m,l}) \\ &= (\boldsymbol{k}_j^{m,l}/\|\boldsymbol{k}_j^{m,l}\|)^\top (\boldsymbol{q}_i^{m,l}/\|\boldsymbol{q}_i^{m,l}\|), \end{aligned} \quad (1)$$

where $\boldsymbol{q}_j^{m,l} \in \boldsymbol{Q}_{m,l}$, $\boldsymbol{k}_j^{m,l} \in \boldsymbol{K}_{m,l}$. Besides, we also consider the attention coefficient $\boldsymbol{B}_{m,l-1}^{KQ}$ from the last LA as the cascade indicator, which reserves historical information from previous LA memories. Hence, we can calculate the attention coefficient using Gated Network (GN) by the combination of the candidate and previous memories, i.e., $\boldsymbol{B}_{m,l}^{KQ} = \mathrm{GN}(\tilde{\boldsymbol{B}}_{m,l}^{KQ}, \boldsymbol{B}_{m,l-1}^{KQ})$, which is computed as:

$$\boldsymbol{G}_{m,l}^{KQ} = \sigma(\boldsymbol{W}_{m,l}^{KQ}[\tilde{\boldsymbol{B}}_{m,l}^{KQ}; \boldsymbol{B}_{m,l-1}] + \boldsymbol{b}_{m,l}^{KQ}), \quad (2)$$

$$\boldsymbol{B}_{m,l}^{KQ} = \boldsymbol{G}_{m,l}^{KQ} \circ \tilde{\boldsymbol{B}}_{m,l}^{KQ} + (1 - \boldsymbol{G}_{m,l}^{KQ}) \circ \boldsymbol{B}_{m,l-1}^{KQ}, \quad (3)$$

where $[;]$ is concatenation operator, σ is sigmoid activation, $\circ$ is Hadamard product. $\boldsymbol{Q}_{m,l}$ is updated as $\boldsymbol{Q}'_{m,l} = \boldsymbol{K}_{m,l}\boldsymbol{A}_{m,l}^{KQ}$. As the output of LA, $\boldsymbol{Q}'_l$ is concatenation of $\{\boldsymbol{Q}'_{m,l}\}$. Based on Basic LA, we develop THREE *LAs:

1) **Bi-directional Linear Attention (BLA)** is denoted as $\boldsymbol{Q}', \boldsymbol{K}' = \mathrm{BLA}(\boldsymbol{Q}, \boldsymbol{K})$, which aims to update $\boldsymbol{Q}$ and $\boldsymbol{K}$ by attention weights $\boldsymbol{A}_m^{KQ} = \alpha + \boldsymbol{B}_{m,l}^{KQ}$, $\boldsymbol{A}_m^{QK} = \alpha + \boldsymbol{B}_{m,l}^{QK}$, where the later coefficient $\boldsymbol{B}_{m,l}^{QK}$ is computed as $\boldsymbol{B}_{m,l}^{QK} = \mathrm{GN}(\tilde{\boldsymbol{B}}_{m,l}^{QK}, \boldsymbol{B}_{m,l-1}^{QK})$, whose candidate state is:

$$\begin{aligned}\tilde{\boldsymbol{B}}_{m,l}^{QK}(j,i) &= \beta(\boldsymbol{q}_j^{m,l}, \boldsymbol{k}_i^{m,l}) \\ &= (\boldsymbol{q}_j^{m,l}/\|\boldsymbol{q}_j^{m,l}\|)^\top (\boldsymbol{k}_i^{m,l}/\|\boldsymbol{k}_i^{m,l}\|), \end{aligned} \quad (4)$$

and the output is $\boldsymbol{Q}'_{m,l} = \boldsymbol{K}_{m,l}\boldsymbol{A}_{m,l}^{KQ}$, $\boldsymbol{K}'_{m,l} = \boldsymbol{Q}_{m,l}\boldsymbol{A}_{m,l}^{QK}$.

2) **Signed Linear Attention (SLA)** is represented as $\boldsymbol{Q}', \boldsymbol{K}' = \mathrm{SLA}(\boldsymbol{Q}, \boldsymbol{K})$, where SLA is integrated with channels of both positive (p) and negative (n) attention coefficients to capture positive and negative relations of input. Therefore, $\pm\boldsymbol{B}_{m,l}$ are employed to gain attention weights:

$$\boldsymbol{A}_{m,l}^{KQ,\mathrm{p}} = \alpha + \boldsymbol{B}_{m,l}^{KQ,\mathrm{p}}, \boldsymbol{A}_{m,l}^{KQ,\mathrm{n}} = \alpha - \boldsymbol{B}_{m,l}^{KQ,\mathrm{n}}, \quad (5)$$

$$\boldsymbol{A}_{m,l}^{QK,\mathrm{p}} = \alpha + \boldsymbol{B}_{m,l}^{QK,\mathrm{p}}, \boldsymbol{A}_{m,l}^{QK,\mathrm{n}} = \alpha - \boldsymbol{B}_{m,l}^{QK,\mathrm{n}}, \quad (6)$$

where $\boldsymbol{B}_{m,l}$ are computed by candidate (Eqs. 1, 4) and previous memories of coefficient using GN (Eqs. 2, 3). Then $\boldsymbol{Q}_l$ and $\boldsymbol{K}_l$ are updated as:

$$\boldsymbol{Q}_{m,l}^{\mathrm{p}} = \boldsymbol{K}_{m,l}\boldsymbol{A}_{m,l}^{KQ,\mathrm{p}}, \boldsymbol{Q}_{m,l}^{\mathrm{n}} = \boldsymbol{K}_{m,l}\boldsymbol{A}_{m,l}^{KQ,\mathrm{n}}, \quad (7)$$

$$\boldsymbol{K}_{m,l}^{\mathrm{p}} = \boldsymbol{Q}_{m,l}\boldsymbol{A}_{m,l}^{QK,\mathrm{p}}, \boldsymbol{K}_{m,l}^{\mathrm{n}} = \boldsymbol{Q}_{m,l}\boldsymbol{A}_{m,l}^{QK,\mathrm{n}}, \quad (8)$$

and the output of query and key is processed as:

$$\boldsymbol{Q}'_{m,l} = \boldsymbol{W}[\boldsymbol{Q}_{m,l}^{\mathrm{p}}; \boldsymbol{Q}_{m,l}^{\mathrm{n}}] + \boldsymbol{b}, \quad (9)$$

$$\boldsymbol{K}'_{m,l} = \boldsymbol{W}[\boldsymbol{K}_{m,l}^{\mathrm{p}}; \boldsymbol{K}_{m,l}^{\mathrm{n}}] + \boldsymbol{b}. \quad (10)$$

3) **Recurrent Linear Attention (RLA)** is a variant of RNN with GRU as the unit, and is written as $\boldsymbol{X}' = \mathrm{RLA}(\boldsymbol{X})$. To study useful semantics from the output of last RLA, each $\boldsymbol{x}_{t,l} \in \boldsymbol{X}_l$ is updated by the memory of the last RLA as $\tilde{\boldsymbol{x}}_{t,l} = \mathrm{LA}(\boldsymbol{x}_{t,l}, \boldsymbol{X}_{l-1})$, where $\boldsymbol{X}_{l-1}$ is used as a cascade indicator. We exploit one GRU layer to balance semantics of the current and previous memories: $\boldsymbol{m}_{t,l} = \mathrm{GRU}(\boldsymbol{x}_{t,l}, \tilde{\boldsymbol{x}}_{t,l})$. Then we introduce GRU network to compute hidden states for each $\boldsymbol{x}_{t,l}$ iteratively. Finally, we study forward & backward sequential features ($\overrightarrow{\boldsymbol{X}} = \overrightarrow{\mathrm{RNN}_{\mathrm{GRU}}}(\boldsymbol{X})$ and $\overleftarrow{\boldsymbol{X}} = \overleftarrow{\mathrm{RNN}_{\mathrm{GRU}}}(\boldsymbol{X})$) that are concatenated into the output $\boldsymbol{X}' = \boldsymbol{W}_{\mathrm{RLA}}[\overrightarrow{\boldsymbol{X}}; \overleftarrow{\boldsymbol{X}}] + \boldsymbol{b}_{\mathrm{RLA}}$.

3.2 INPUT LAYER

BERT is the token-level initializer for input sequence $\boldsymbol{I} = \mathrm{BERT}([\mathtt{CLS}]\mathbb{C}[\mathtt{SEP}]\mathbb{S}_i[\mathtt{SEP}])$. We split $\boldsymbol{I}$ into $\boldsymbol{C}_0$ and $\boldsymbol{S}_0$ for the negative cue $\mathbb{C}$ and the i-th sentence $\mathbb{S}_i$, respectively, and $\boldsymbol{C}_0$ is collapsed into a vector $\boldsymbol{c}_0$. Thus, we consider the upgrade of dual-channel memories: the token-level memory $\boldsymbol{c}_l$ and sentence-level memory $\boldsymbol{S}_l$.

3.3 CASCADE LINEAR ATTENTION LAYER

This layer consists of L identical **CLAN units** computed as $\boldsymbol{c}_{l+1}, \boldsymbol{S}_{l+1} = \Psi_l(\boldsymbol{c}_l, \boldsymbol{S}_l)$, where Ψ_l is the l-th unit, and $l \in [0, L)$. To learn the representation of cue, $\boldsymbol{c}_l$ is concatenated with each token $\boldsymbol{s}_l \in \boldsymbol{S}_l$ to form another channel of the sentence $\boldsymbol{T}_l = \mathrm{FC}([\boldsymbol{C}_l; \boldsymbol{S}_l])$, where $\boldsymbol{C}_l = \{\boldsymbol{c}_l\}_{i=0}^{|\mathbb{S}_i|-1}$, and FC is a fully-connected layer. Then $\boldsymbol{S}_l$ and $\boldsymbol{T}_l$ are fed into *LAs below. Concretely, each Ψ_l is made up with several networks based on the following **Human-like Reading Strategies** (**HRS**):

HRS1: *Sentence-level Global Reading*. This strategy is applied when reading sentences globally and quickly for

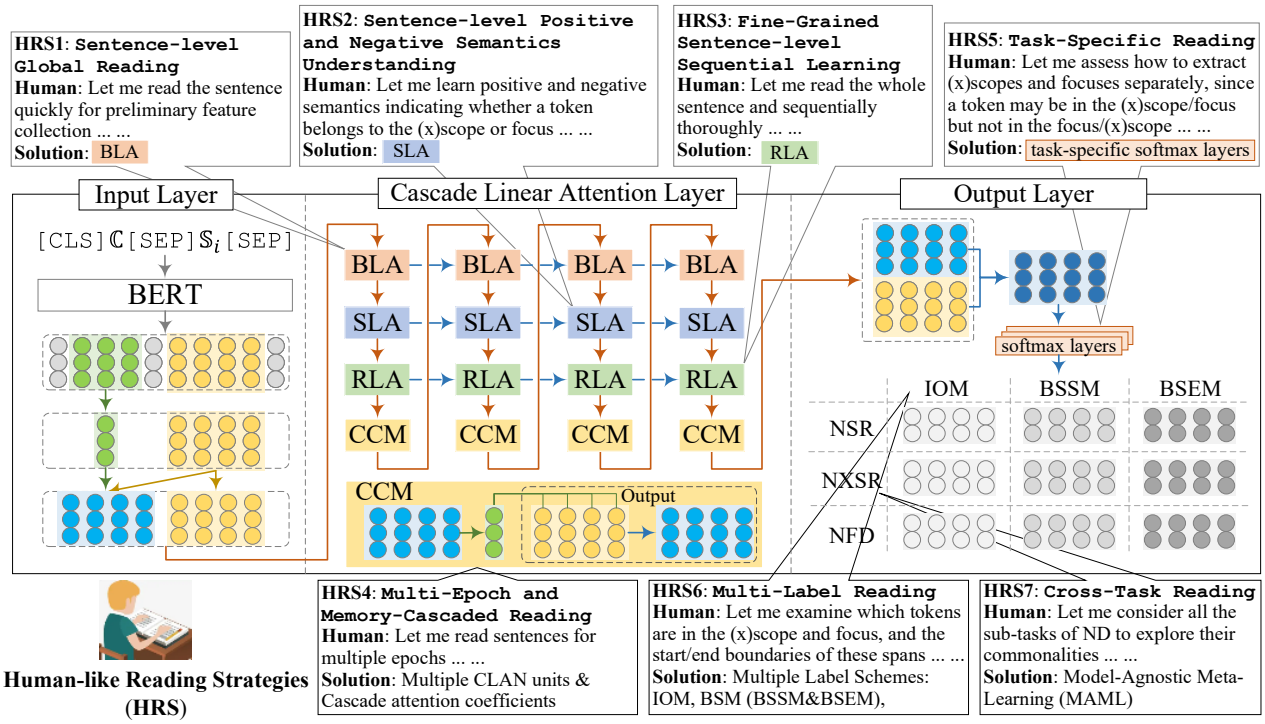


Figure 2: The overall architecture of Cascade Linear Attention Network (CLAN) model.

preliminary feature collection, and implemented by BLA: $\boldsymbol{S}_l^{(1)}, \boldsymbol{T}_l^{(1)} = \mathrm{BLA}(\boldsymbol{S}_l, \boldsymbol{T}_l | \boldsymbol{B}_{l-1}^*)$, where $\boldsymbol{B}_{l-1}^*$ means we apply cascade attentions in which attention coefficients are updated by those from the last corresponding *LA.

HRS2: *Sentence-level Positive and Negative Semantics Understanding*. We introduce this strategy because humans tend to capture positive and negative semantics of texts to determine which texts are negative statements and which are not during reading. Specifically, humans will learn purposefully whether a token belongs to the negative (x)scope or focus for ND. Hence, HRS2 is accomplished by SLA with signed coefficients: $\boldsymbol{S}_l^{(2)}, \boldsymbol{T}_l^{(2)} = \mathrm{SLA}(\boldsymbol{S}_l^{(1)}, \boldsymbol{T}_l^{(1)} | \boldsymbol{B}_{l-1}^*)$.

HRS3: *Fine-Grained Sentence-level Sequential Learning*. Eventually, humans have a strong disposition to read the whole sentence thoroughly again, and therefore can learn the sentence-level sequential semantics before making the final decision. This process can be simulated by RLA: $\boldsymbol{S}_l^{(3)} = \mathrm{RLA}_l(\boldsymbol{S}_l^{(2)})$, $\boldsymbol{T}_l^{(3)} = \mathrm{RLA}_l(\boldsymbol{T}_l^{(2)})$.

HRS4: *Multi-Epoch and Memory-Cascaded Reading*. Humans usually read sentences for multiple epochs when texts are quite complicated, leading to the difficulty in detecting (x)scopes & focuses, and then understand the task of extracting negative statements gradually based on those cascade memories from the last epoch. We thereby adhibit attention coefficients from the last unit as the cascade indicator for each *LA. Furthermore, we set a **Collapse and Concatenation Module (CCM)** to process the output of RLA (output of HRS3): $\boldsymbol{T}_{l+1}$(i.e., $\boldsymbol{T}_{l+1}$=$\boldsymbol{T}_l^{(3)}$) is compressed into $\boldsymbol{c}_{l+1}$, which is the input for the next unit Ψ_{l+1} along with $\boldsymbol{S}_{l+1}$(i.e., $\boldsymbol{S}_{l+1}$=$\boldsymbol{S}_l^{(3)}$). After L units, memories of the cue and sentence are updated to $\boldsymbol{c}_L$ and $\boldsymbol{S}_L$, respectively.

3.4 OUTPUT LAYER

The final unified representation of the sentence is $\boldsymbol{H}_{\mathbb{S}} = \mathrm{FC}([\boldsymbol{C}_L; \boldsymbol{S}_L])$, where $\boldsymbol{C}_L$=$\{\boldsymbol{c}_L\}_{i=0}^{|\mathbb{S}_i|-1}$. Next, we compute probability distributions of labels of tokens for each sub-task of negation detection according to more HRSs as follows:

HRS5: *Task-Specific Reading*. Humans assess how to extract (x)scopes and focuses separately since a token may be in the (x)scope/focus but not in the focus/(x)scope. Hence, we feed $\boldsymbol{H}_{\mathbb{S}}$ into the corresponding task-specific softmax layers $\boldsymbol{O}_{\tau,\eta} = \mathrm{softmax}_{\tau,\eta}(\mathrm{FC}_{\tau,\eta}(\boldsymbol{H}_{\mathbb{S}}))$, where the sub-task $\tau \in \{\mathrm{NSR}, \mathrm{NXSR}, \mathrm{NFD}\}$, and η denotes label schemes inspired by a novel HRS as below:

HRS6: *Multi-Label Reading*. It aims to hedge spans from a multi-granularity view, because humans not only examine which tokens are in the (x)scope and focus, but pay more attentions to start and end boundaries of spans. Accordingly, our label scheme is assembled by two map frameworks:

1) *IO Map* (*IOM*), where a token is labelled as I/O if it is IN/OUTSIDE(NOT IN) the (x)scope or focus;

2) *Boundary Shift Map* (*BSM*) [Wu and Sun, 2023] includes

Algorithm 1 Optimization of CLAN model

Input: Training set, in which the i-th sample is denoted as $\{\mathbb{C}_i, \mathbb{S}_i, \{y^{(i)}_{\tau,\eta}\}\}$, where $\mathbb{C}$ is the cue, $\mathbb{S}_i$ is the sentence, $\{y^{(i)}_{\tau,\eta}\}$ is set of annotated labels, the sub-task $\tau \in$ {NSR, NXSR, NFD}, label scheme $\eta \in$ {IOM, BSSM, BSEM}.
Output: Trained CLAN.

1: **Phase 1**: Execute warm-up multi-task learning for all sub-tasks.
2: **Phase 2**: Model-Agnostic Meta-Learning:
3: **for** $j \in$ **do**
4: Sample the j-th batch $\mathcal{B}_j$;
5: **for** each task-task τ **do**
6: Optimize parameters on the support set $\widetilde{\Theta} = \Theta - \alpha\nabla_\Theta \mathcal{L}_\tau(\mathcal{B}_j^{\text{sup}}, \Theta)$, where α is the learning rate;
7: **end for**
8: Update the parameters on the query set: $\Theta = \Theta - \beta\nabla_\Theta \sum_\tau \mathcal{L}_\tau(\mathcal{B}_j^{\text{qry}}, \widetilde{\Theta})$, where β is the learning rate;
9: **end for**
10: **Phase 3**: Post multi-task learning for all sub-tasks.

Boundary Shift Start Map (BSSM) & Boundary Shift End Map (BSEM) that tag start/end tokens as Bs/Be, and annotates each token as L/R if the nearest boundary resides on the Left/Right of this token.

To sum up, the loss function of each task τ is designed as:

$$\mathcal{L}_\tau = \sum_\eta \epsilon_{\tau,\eta}\mathcal{L}_{\tau,\eta},\ \mathcal{L}_{\tau,\eta} = \frac{1}{N}\sum_{i=0}^{N-1} \log p(y^{(i)}_{\tau,\eta}|\Theta), \quad (11)$$

where $\eta \in$ {IOM, BSSM, BSEM}, $y^{(i)}_{\tau,\eta}$ is golden label, N is the number of samples. The complete loss function of CLAN model using Multi-Task Learning (MTL) is $\mathcal{L}_{\text{CLAN}} = \sum_\tau \gamma_\tau \mathcal{L}_\tau$, where ϵ and γ are trade-off coefficients.

HRS7: *Cross-Task Reading*. Finally, humans can take into account all the sub-tasks of ND to explore their commonalities, instead of being limited to each task. According to this strategy, Model-Agnostic Meta Learning (MAML) is led into our model, and the three-stage optimization process is displayed in Algorithm 1: 1) Train all the sub-tasks by warm-up MTL; 2) Compute the adapted parameters $\widetilde{\Theta}$ on the support set for each task, and then update the model by evaluating $\widetilde{\Theta}$ on the query set for all tasks; 3) Fine-tune all the sub-tasks by post MTL.

4 EXPERIMENTATION

4.1 EXPERIMENTAL SETTINGS

Corpus and Evaluation Details We employ SOCC [Kolhatkar et al., 2020], PB-FOC [Morante and Blanco, 2012], and CNeSp [Sheng et al., 2020] as benchmark datasets. SOCC and CNeSp annotate negative cues, scopes, focuses, while SOCC gives xscopes. PB-FOC is a widely-accepted focus corpus without scopes, and we extract scopes by a variant model of [Fancellu et al., 2016], just as [Hossain et al., 2020]. For more details of statistics of SOCC, PB-FOC, CNeSp, please refer to Table 2, 3, 4 in Appendix B.1. We use annotated cues directly. As previous work, we implement 10-fold cross-validation on SOCC, CFA, CPR, CSL, and report the performance on the test set of PB-FOC. For evaluation metrics, we employ the ACCuracy (ACC) of predicted samples that are exact match with the annotated samples as the strict criterion.

Baselines of NSR/NXSR Neural network baselines are classified into the following categories: 1) **BLSTM** [Fancellu et al., 2016] is based on Bi-directional LSTM with lexical features; 2) **X-YP** formulates NSR/NXSR as token classification tasks, and selects syntactic Path between the cue and each token in parse trees. X denotes CNN [Qian et al., 2016] and CBLSTM [Lazib et al., 2020] (CNN+BLSTM). Y$\in$ {C, D} represents Constituency and Dependency parse trees; 3) **RvNN-Y** is Recursive Neural Network working on syntactic parse trees. Similar to X-YP models above, Y can also be assigned as C when using Constituency trees [McKenna and Steedman, 2020], or D for Dependency trees [Fei et al., 2020]; 4) **RNet** [Qian et al., 2023] is the pipeline of BERT and Residual Networks; 5) **BA-GNN** [Iyer et al., 2021] is Bi-level Attention Graph Neural Network that models node-&graph-level attention by capturing hierarchical dependencies and contextual interactions, and learns representations over explicit graph structures.

Baselines of NFD Relative models cover two groups: 1) **WTG Models** are based on Word-based and Topic-driven Graph structure, including Supervised version **SWTG** [Zou et al., 2014] and Unsupervised version **UWTG** [Zou et al., 2015]; 2) **BA-GNN** [Iyer et al., 2021] models both node-node and relation-relation interactions; 3) **BERT+LSTM+CRF Models** exploit BERT for word embeddings, and adopt LSTM and CRF to capture sequential information: 3.1) **LSTM-C** is made up of BERT+LSTM+CRF; 3.2) **LSTM-TA** [Shen et al., 2019] integrates with topic-level contextual attention; 3.3) **LSTM-ML** [Hossain et al., 2020] considers lexical, syntactic, and semantic features for a multi-layer LSTM model.

4.2 OVERALL RESULTS AND ANALYSIS

Table 1 exhibits performance of CLAN and baselines on N(X)SR, NFD. Detailed analysis is discussed as follows:

Performance on NSR. Compared to the simple BLSTM, other CNN/LSTM/RvNN models employ parse trees as the primary syntactic features that can construct correlations among remote chunks within scopes, and can outperform BLSTM. Although adopting BERT as the backbone encoder,

RNet still cannot defeat our model, manifesting the downstream task-specific module is too simple to learn sufficient semantics for scopes. BA-GNN attempts to learn adequate information from interactions of tokens and their relations via coarse-grained graph networks by extracting semantics just from local neighborhood contexts. Besides, BA-GNN is a general model, rather than a task-specific model designed for negation detection. Therefore, BA-GNN cannot capture high-level effective information among negative cues, (x)scopes, focuses. Our CLAN model can achieve the best performance, and the reasons are three-fold: 1) CLAN introduces the complex architecture with heterogeneous linear attentions; 2) We employ two types of label schemes for the diversity of the representation of scopes; 3) We can capture extra common information from focuses by meta-learning, since focuses are cores of negation and fundamental elements that constitutes scopes.

Performance of NFD. Our model can beat both topic-driven graph and BERT+LSTM+CRF families. The graph structures of tokens learned by WTG and BA-GNN are quite simple and shallow, and CLAN can capture interactions among tokens and cues at the level of multi-granularity with *LAs. Compared with WGT & BA-GNN, BERT+LSTM+CRF can extract sequential dependencies of tokens, but cannot surpass CLAN. Owing to HRSs, our *LA-based model can not only encode token-level and sequence-level semantics, but benefit from NSR that can narrow the range of candidate texts for focuses.

Performance of NXSR. NXSR gets the lowest results among three sub-tasks due to the minority of samples. Although the definition of xscopes is similar to scopes, cues are usually not in xscope, leading to remote relationship and increase in the difficulty of detection. In addition, NFD does not bring significant profits for NXSR, since focuses often rely heavily on scopes. CLAN still outperforms other baselines, which can validate its advantages and the benefit from NSR.

4.3 ABLATION STUDY

Performance of ablation of CLAN is shown in Figure 3.

Linear Attentions. Figure 3a shows the performance of the model will be confronted with varying degrees of decline when *LAs are abandoned, verifying that *LAs are all useful encoders of CLAN model. According to the margin of the decrease of ACC, BLA is not so significant as SLA and RLA, since both SLA and RLA are integrated with bi-directional encoding in BLA, and collect more features (positive/negative attention coefficients, sequential semantics) than BLA. Furthermore, RLA is the most meaningful *LA, especially for NFD, because the detection of (x)scopes and focuses rely heavily on sequential relations of tokens.

Learning Frameworks. According to Figure 3b, if we drop

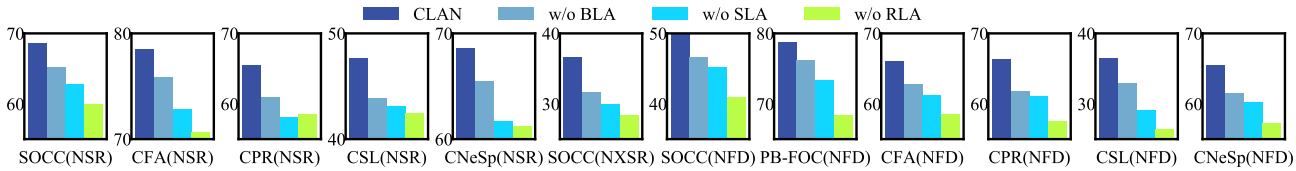
Table 1: Performance of models on N(X)SR and NFD. For focuses, "Sgl/Cue/Oth" means the focus is *a single token but not negative cue*, *a negative cue*, or *belongs to neither Sgl nor Cue*.

(a) NSR and NXSR

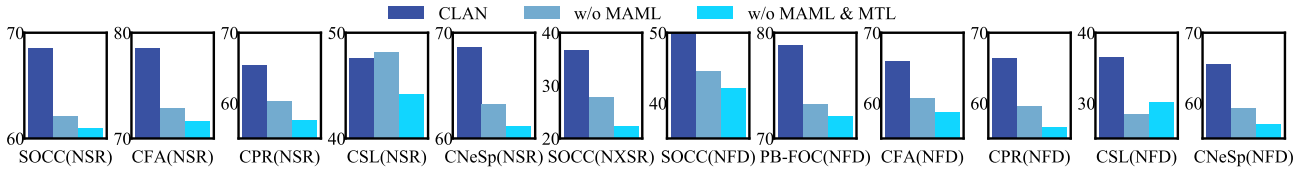
Models	NSR					NXSR
	SOCC	CNeSp				SOCC
		CFA	CPR	CSL	ALL	
BLSTM	51.57	65.69	49.77	36.92	53.51	21.83
CNN-CP	55.82	70.17	53.65	40.68	57.66	25.33
CNN-DP	57.60	71.23	53.31	42.59	57.40	27.83
CBLSTM-CP	57.29	70.54	55.16	43.03	59.23	26.50
CBLSTM-DP	59.33	71.08	56.92	44.70	60.52	28.50
RvNN-C	60.15	72.75	55.84	44.27	59.78	29.83
RvNN-D	63.48	74.46	60.03	48.15	64.29	32.17
RNet	64.71	74.14	58.29	**48.34**	62.83	31.50
BA-GNN	61.42	71.30	54.58	43.53	59.11	30.17
CLAN	**68.56**	**78.51**	**65.42**	47.61	**68.58**	**36.67**

(b) NFD

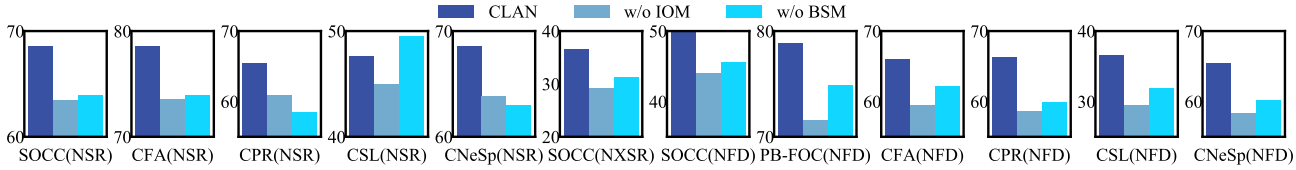
Models	SOCC	PB-FOC	CNeSp			
			CFA	CPR	CSL	ALL
WTG	36.81	67.13	53.88	54.67	24.20	53.58
Sgl	37.63	66.67	30.50	57.81	0.00	54.20
Cue	-	65.41	-	-	-	-
Oth	33.50	67.98	55.66	53.86	25.42	53.46
UWTG	38.74	69.38	56.27	56.68	28.39	55.79
Sgl	41.66	70.83	35.58	60.45	20.00	57.36
Cue	-	67.03	-	-	-	-
Oth	36.49	70.07	57.85	55.71	25.61	55.48
BA-GNN	36.23	65.45	55.92	55.11	25.89	54.52
Sgl	36.84	65.63	32.17	56.34	0.00	53.14
Cue	-	62.16	-	-	-	-
Oth	35.59	66.82	57.87	54.79	27.17	54.78
LSTM-C	40.29	67.56	59.23	60.41	31.38	59.29
Sgl	44.17	68.75	39.15	63.90	10.00	60.68
Cue	-	65.41	-	-	-	-
Oth	36.05	68.21	60.76	59.52	29.45	59.03
LSTM-TA	43.96	70.51	60.94	62.08	29.70	60.87
Sgl	46.43	71.88	41.73	65.97	10.00	62.82
Cue	-	69.19	-	-	-	-
Oth	40.07	70.77	62.41	61.19	27.70	60.50
LSTM-ML	46.19	75.70	62.95	63.04	33.31	62.22
Sgl	51.82	79.17	41.80	67.93	**40.00**	64.85
Cue	-	73.51	-	-	-	-
Oth	42.54	75.87	64.56	61.78	29.54	61.71
CLAN	**50.38**	**78.79**	**65.98**	**66.36**	**36.51**	**65.46**
Sgl	**54.75**	**81.25**	**46.06**	**70.50**	30.00	**67.51**
Cue	-	**77.84**	-	-	-	-
Oth	**45.61**	**78.65**	**67.50**	**65.29**	**33.55**	**64.97**



(a) Ablation of linear attentions.



(b) Ablation of learning frameworks.



(c) Ablation of label schemes.

Figure 3: Performance of ablation study of CLAN model.

MAML and only use Multi-Task Learning (MTL) to train ND sub-tasks jointly, we can observe decrease in each sub-task, especially NFD, demonstrating that by pure MTL, scopes (and xscopes in SOCC) are likely to restrain our model from learning more meaningful information of focuses, resulting in training imbalance. The main reason is that (x)scopes may cause the model to tend to extract all negative texts, rather than the most dominant ones (i.e., focuses) that are just a few tokens in scopes. This issue are addressed by MAML that can fine-tune each task on the support set. Moreover, if we discard MAML & MTL, and train each sub-task in isolation, we will get even worse results than using MAML or MTL, since we cannot learn semantics from other sub-tasks or boost a sub-task by each other.

Label Scheme. Figure 3c can prove both IOM and BSM are useful schemes to tag tokens, since removing IOM/BSM leads to declines in performance of ND task. BSM is quite an efficacious supplementary paradigm, because start and end indexes are valuable clues and can delimit focuses and (x)scopes clearly. On most datasets, IOM is more meaningful than BSM. The principal reason is that tags of start and end tokens (Bs/Be) are much fewer than other labels indicating tokens in/not in the negative texts. But there are also exceptions, i.e., CPR, on which removing BSM leads to more decline in performance of NSR than removing IOM. This result can be attributed to the much shorter lengths of scopes. Then, the scarcity of Bs/Be is greatly alleviated, resulting in hedging scopes by start and end token is easier ◦

Cross-Domain Detection. To evaluate generalization ability of CLAN, we get cross-domain performance by treating CNeSp as one Chinese dataset and SOCC&PB-FOC as one English dataset, and present the results in Figure 4. We observe that our model fails to surpass the in-domain version where trained on CFA, CPR, CSL, SOCC, PB-FOC separately. The major reason is we cannot extract more beneficial features from (x)scopes & focuses, because they do not share high-level semantics across different texts.

4.4 CASE STUDY

Figure 5 illustrates that CLAN is able to implement mutual promotion for ND sub-tasks, and is beneficial to several samples that are predicted correctly:

NSR. In S2, the annotated scope of "*not*" is the whole clause until the end of the sentence. If we just apply CLAN to NSR, the predict scope neglects the last several words, among which "*power and dominance*" is the golden focus according to the positive text "*mutual defense*". CLAN model detects the focus successfully, and identifies it as part of the scope via MAML mechanism;

NFD. In S3, the negative focus of the cue "*hasn't*" is "*the Canadian economy*", because the positive statement is "*every other G7 economy*". We may obtain the wrong focus "*economy*" if we solely consider NFD in CLAN. By bringing in MAML for both NSR and NFD, we are able to hedge the correct scope consisting of two parts, i.e., the nominal phrase "*the skill, intelligence or aptitude*" and the verbal phrase "*manage the Canadian economy*", which guides our model to detect the correct focus;

NXSR. S4 can exemplify that NXSR is beneficial from other tasks as well. The incomplete xscope "*failing Syrian*

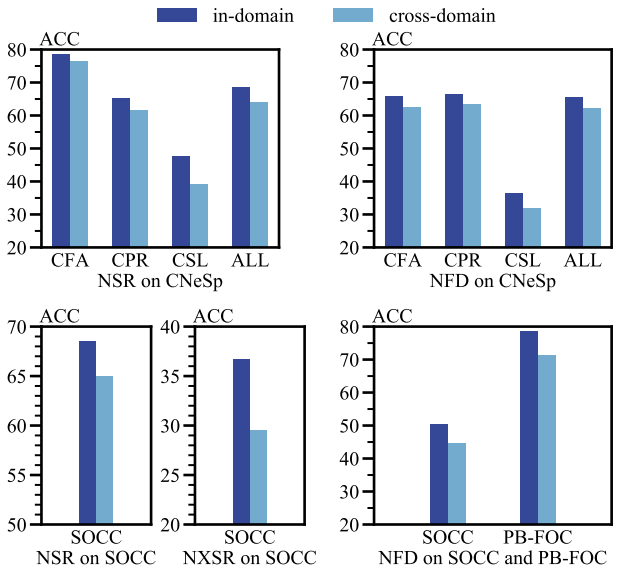


Figure 4: Performance of CLAN model on monolingual in-domain and cross-domain evaluation.

S2: *Canada and U.S. are allies in mutual defense , but mutual defense does not mean we get dragged into global wars of power and dominance ...*
Scope predicted by CLAN-NSR: *not mean we get dragged into global wars*

S3: *We have negative GDP growth, a plunging dollar, a negative trade balance, worsening unemployment and gains in productivity are trailing every other G7 economy . He hasn't the skill, intelligence or aptitude to manage the Canadian economy and he never did.*
Focus predicted by CLAN-NFD: *economy*

S4: *" Europe may be failing Syrian refugees " is Nonsense, and a more accurate statement is that the culture of sectarian violence, misogyny, and intolerance have failed the residents of Arabic nations, including Syria.*
XScope predicted by CLAN-NXSR: *Syrian refugees*

Figure 5: Several cases predicted by CLAN correctly, where CLAN-NSR/NFD/NXSR denotes CLAN is trained only for NSR/NFD/NXSR, respectively.

refugees" is obtained if we abandon meta-learning and perform NXSR as an isolated task. This result can be rectified by MAML, and we can extract the correct focus "*Europe*", which helps CLAN to be conscious that "*Europe*" belongs to the xscope, and then give the right result.

4.5 ERROR ANALYSIS

Several typical error cases are given in Figure 6, which includes those mistakenly predicted focuses of NFD (S5, S6, S7), scopes of NSR (S8, S9), xscopes of NXSR (S10), and can indicate that the performance of our model can be improved by revising them in the future.

S5: *So if you did not opt out of Uber's arbitration clause within 30 days of accepting the agreement ..., please CONTACT US so that we can add you to our list of drivers to pursue individual arbitration ... If you accepted an Uber contract within the last 30 days , you can still opt out of the arbitration clause by emailing optout@uber.com ...*
Prediction: *accepting the agreement*

S6: *People of America want a president who actually says what every common man is thinking but no one listens .*
Prediction: *no one listens*

S7: *Mr. Paul had no right to buy art for the S&L in the first place – it is not on the "permissible" list – without seeking a special dispensation, which he did not do.*
Prediction: *seeking a special dispensation*

S8: 我们要求看房间时，服务员很不情愿地给了我们一张房卡。(*When we requested to see the room, although the waiter is not willing to, he still gave us a room card.*)
Prediction: 服务员很不情愿地给了我们一张房卡(*although the waiter is not willing to, he still gave us a room card*)

S9: 如此则会导致密文数据查询不再具有明文查询特性而造成安全性方面的挑战，且云环境下密文查询有以下需求...(*This will result in the fact that cipher data query no longer has features of clear data query that leads to security challenges, and there are following requirements for cipher data query in the cloud environment:...*)
Prediction: 密文数据查询不再具有明文查询特性而造成安全性(*cipher data query no longer has features of clear data query that leads to security*)

S10: *Indeed, if Harper can control the economy then why did we have a downturn? Truth is he can't.*
Prediction: *control the economy then why did we*

Figure 6: Several error cases predicted by CLAN model.

Focuses. Errors are analyzed as follows: 1) *Positive statements*. NFD mainly depends on semantic understanding of positive texts, which may become more difficult when the sentence is long and complex. In S5, focus should be "*within 30 days of accepting the agreement*" according to the positive argument "*accepted an Uber contract within the last 30 days*", but our model only detects the verbal phrase "*accepting the agreement*" and omits the temporal adverbial. 2) *Single-word focuses*. Compared with (x)scopes, single-word focuses occupy a considerable proportion, but faces word-level sparsity. In S6, the focus of "*not*" should be "*listens*", since the positive statement is "*says*". But our

model predicts the scope "*no one listens*" as the focus. Furthermore, PB-FOC annotates a negative cue as a focus when no verbal action occurred, nor verbs carry positive meanings. It can be observed that all clauses of S7 are negated by negative cues "*no*, *not*, *without*, *not*" sequentially, and then the focus of the last "*not*" is itself due to the lack of affirmative texts. Our model gives a wrong result "*seeking a special dispensation*", which is actually the xscope of "*not*".

Scopes. In general, scopes usually are clauses or chunks with complete syntactic structures. However, NSR task may become harder when encountering scopes with complicated structures. For S8, "*not*" only negates "*willing*" instead of "*gave*", since "*waiter is not willing*" is a manner adverbial of the verb "*gave*". Without capturing these semantics, our model negates the whole clause. Similarly, S9 is quite long, but "*no long*" only acts on "*has features of clear data query*", rather than the complete current clause. These cases prove the significance and difficulty of extracting negation.

XScope. The main problem of NXSR is the insufficiency of samples, despite of advanced techniques (e.g., MAML, multiple schemes of labels). Besides, the similar definitions of scopes and xscope also lead to the inadequate training for NXSR, since scopes are much more than xscopes. S10 shows CLAN predicts an incorrect end boundary due to the failure of detecting the structure of "*if ... then ...*", although realizing "*can't*" should negate a verb phrase.

5 CONCLUSION

This paper is dedicated to Negation Detection (ND). To solve the problem of unified modelling for ND sub-tasks, we propose a unified framework that introduces Model-Agnostic Meta-Learning (MAML) to implement mutual promotion for ND sub-task, including Negative Scope Resolution (NSR), Negative Focus Detection (NFD), Negative XScope Resolution (NXSR). In terms of fine-grained encoder, based on human reading mechanism, we design Cascade Linear Attention Network (CLAN) by heterogeneous linear attentions with cascade coefficients. Experimental results on several widely-acceptable datasets can prove the usefulness of our model.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62376181, 62276177, 62006167), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Nos. 25KJA521001 and 24KJB520036), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Eduardo Blanco and Dan I. Moldovan. Fine-grained focus for pinpointing positive implicit meaning from negated statements. In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings (NAACL 2012)*, pages 456–465, 2012.

Ramona Christen, Anastassia Shaitarova, Matthias Stürmer, and Joel Niklaus. Resolving legalese: A multilingual exploration of negation scope resolution in legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC/COLING 2024)*, pages 13992–14004, 2024.

Federico Fancellu, Adam Lopez, and Bonnie L. Webber. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

Federico Fancellu, Adam Lopez, Bonnie L. Webber, and Hangfeng He. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 58–63, 2017.

Hao Fei, Yafeng Ren, and Donghong Ji. Negation and speculation scope detection using recursive neural conditional random fields. *Neurocomputing*, 374:22–29, 2020.

Mareike Hartmann and Anders Søgaard. Multilingual negation scope resolution for clinical text. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis (LOUHI@EACL 2021)*, pages 7–18, 2021.

Md Mosharaf Hossain, Kathleen E. Hamilton, Alexis Palmer, and Eduardo Blanco. Predicting the focus of negation: Model and error analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 8389–8401, 2020.

Roshni G. Iyer, Wei Wang, and Yizhou Sun. Bi-level attention graph neural networks. In *IEEE International Conference on Data Mining, ICDM 2021*, pages 1126–1131. IEEE, 2021.

Aditya Khandelwal and Suraj Sawant. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of The 12th Language Resources and Evaluation Conference, (LREC 2020)*, pages 5739–5748, 2020.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4:155–190, 2020.

Natalia Konstantinova, Sheila C. M. de Sousa, Noa P. Cruz Díaz, Manuel J. Maña López, Maite Taboada, and Ruslan Mitkov. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3190–3195, 2012.

Lydia Lazib, Bing Qin, Yanyan Zhao, Weinan Zhang, and Ting Liu. A syntactic path-based hybrid neural network for negation scope detection. *Frontiers of Computer Science*, 14(1):84–94, 2020.

Hao Li and Wei Lu. Learning with structured representations for negation scope extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 533–539, 2018.

Nick McKenna and Mark Steedman. Learning negation scope from syntactic structure. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM@COLING 2020)*, pages 137–142, 2020.

Roser Morante and Eduardo Blanco. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 265–274, 2012.

Roser Morante and Walter Daelemans. ConanDoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1563–1568. European Language Resources Association (ELRA), 2012.

Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo. Speculation and negation scope detection via convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 815–825, 2016.

Zhong Qian, Tiening Sun, Ting Zou, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. Speculation and negation scope resolution via machine reading comprehension formulation with data augmentation. In *Proceedings of the Database Systems for Advanced Applications - 28th International Conference (DASFAA 2023)*, volume 13945, pages 487–496, 2023.

Zhong Qian, Ting Zou, Zihao Zhang, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. Speculation and negation identification via unified machine reading comprehension frameworks with lexical and syntactic data augmentation. *Engineering Applications of Artificial Intelligence*, 131: 107806, 2024.

Sabine Rosenberg and Sabine Bergler. UConcordia: CLaC negation focus detection at *Sem 2012. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 294–300, 2012.

Elena Sergeeva, Henghui Zhu, Amir Tahmasebi, and Peter Szolovits. Neural token representations and negation and speculation scope detection in biomedical and general domain text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI@EMNLP 2019)*, pages 178–187, 2019.

Longxiang Shen, Bowei Zou, Yu Hong, Guodong Zhou, Qiaoming Zhu, and AiTi Aw. Negative focus detection via contextual attention mechanism. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pages 2251–2261, 2019.

Jiaxuan Sheng, Bowei Zou, Longxiang Shen, Jing Ye, and Yu Hong. Research on Chinese negative focus identification: Dataset and baseline. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics (CCL 2020)*, pages 550–560, October 2020.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing 2008 (BioNLP 2008)*, pages 38–45. Association for Computational Linguistics, 2008.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11), 2008.

Yin Wu and Aixin Sun. Negation scope refinement via boundary shift loss. In *Findings of the Association for Computational Linguistics (ACL 2023)*, pages 6090–6099, 2023.

Asahi Yoshida, Yoshihide Kato, and Shigeki Matsubara. Negation scope conversion: Towards a unified negation-annotated dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC/COLING 2024)*, pages 12093–12099, 2024.

Mengyu Zhang, Weiqi Wang, Shuqiao Sun, and Weiwei Sun. Negation scope resolution for chinese as a second language. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications (BEA@EACL)*, pages 1–10, 2021.

Yiyun Zhao and Steven Bethard. How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4729–4747, 2020.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. Negation focus identification with contextual discourse information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 522–530, 2014.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. Unsupervised negation focus identification with word-topic graph model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 1632–1636, 2015.

Reading As Humans: Unified Negation Detection via Cascade Linear Attention Network with Model-Agnostic Meta-Learning (Appendices)

Zhong Qian✉[1] **Peifeng Li**[1] **Qiaoming Zhu**[1] **Guodong Zhou**[1]

[1]School of Computer Science and Technology, Soochow University, Suzhou 215006, China

This appendix discusses more details about the proposed Cascade Linear Attention Network (CLAN) (§ A), experimentation (§ B), and acronyms (§ C).

A MORE DETAILS OF CASCADE LINEAR ATTENTION NETWORK

A.1 LABEL SCHEME

§ 3.4 considers two label schemes that are based on these map frameworks elaborated as follows:

IO Map IOM maps a token into I if this token is in the scope, xscope, focus. Otherwise, IOM maps a token into O if this token is out of the scope, xscope, focus, just as exemplified by the following sentence, where I indicates the scope of the cue "*not*" in S1.1 (Figure 1):

- *Due*/O *to*/O *Mr.*/O *Simpson*/O *'s*/O *health*/O *status*/O *,*/O *the*/I *doctor*/I *does*/I *not*/I *allow*/I *him*/I *to*/I *be*/I *discharged*/I *from*/I *the*/I *hospital*/I *,*/O *but*/O *permits*/O *him*/O *to*/O *take*/O *a*/O *walk*/O *in*/O *the*/O *inpatient*/O *building*/O *.*/O

Boundary Shift Map (BSM) [Wu and Sun, 2023] includes two sub-maps:

- **Boundary Shift Start Map (BSSM)** tags start tokens as Bs, and annotates one token as L/R if the nearest boundary resides on its Left/Right.
- **Boundary Shift End Map (BSEM)** tags end tokens as Be, and annotates one token as L/R if the nearest boundary resides on its Left/Right.

According to BSSM and BSEM, the labels of the scope in the sentences S1.1 are shown as below:

- *Due*/R *to*/R *Mr.*/R *Simpson*/R *'s*/R *health*/R *status*/R *,*/R *the*/Bs *doctor*/L *does*/L *not*/L *allow*/L *him*/L *to*/L *be*/L *discharged*/L *from*/L *the*/L *hospital*/L *,*/L *but*/L *permits*/L *him*/L *to*/L *take*/L *a*/L *walk*/L *in*/L *the*/L *inpatient*/L *building*/L *.*/L
- *Due*/R *to*/R *Mr.*/R *Simpson*/R *'s*/R *health*/R *status*/R *,*/R *the*/R *doctor*/R *does*/R *not*/R *allow*/R *him*/R *to*/R *be*/R *discharged*/R *from*/R *the*/R *hospital*/Be *,*/L *but*/L *permits*/L *him*/L *to*/L *take*/L *a*/L *walk*/L *in*/L *the*/L *inpatient*/L *building*/L *.*/L

B MORE DETAILS OF EXPERIMENTATION

This section gives more details of datasets (§ B.1) and experimental results (§ B.2).

Table 2: Statistics of samples of negation in SOCC corpus. For focuses, "Sgl" means the negative focus is *a single token (or Chinese character, the same below) but not a negative cue*, and "Oth" denotes *longer than one token but not a negative cue*. No negative cue is annotated as a focus. "#Sentences" means the number of sentences containing negative scopes, xscopes, focuses, while "#Samples" denotes the number of samples of scopes, xscopes, or focuses.

	Scope	**XScope**	**Focus**		
			Sgl	**Oth**	**All**
#Sentences	1383	36	776	712	1354
Average Length of Sentences	32.02	35.81	32.86	36.54	32.32
#Samples	1375	34	705	637	1342
Average Length of Samples	6.93	4.12	1.00	3.74	2.30

Table 3: Statistics of samples of focuses in PB-FOC corpus.

		Sgl	**Cue**	**Oth**	**All Focuses**
Train	#Sentences	299	591	1414	2304
	Average Length of Sentences	23.48	26.52	26.79	26.29
	#Samples	299	591	1414	2304
	Average Length of Samples	1.00	1.00	6.81	4.56
Dev	#Sentences	70	137	324	531
	Average Length of Sentences	22.91	28.68	27.42	27.15
	#Samples	70	137	324	531
	Average Length of Samples	1.00	1.00	6.80	4.54
Test	#Sentences	99	182	431	712
	Average Length of Sentences	24.74	26.52	26.90	26.50
	#Samples	99	182	431	712
	Average Length of Samples	1.00	1.00	6.81	4.52

B.1 MORE DETAILS OF DATASETS

B.1.1 SOCC

SFU Opinion and Comments Corpus (SOCC) [Kolhatkar et al., 2020] is a collection of opinion articles and the comments posted in response to the articles that include the opinion pieces published in the Canadian newspaper *The Globe and Mail* between 2012 and 2016. SOCC is part of a project that investigates the linguistic characteristics of online comments, and can be used to study a host of pragmatic phenomena. This corpus presents four layers of annotations, i.e., constructiveness, toxicity, negation and appraisal. In this paper, we mainly pay attention to negation information, whose statistics are shown in Table 2. We exemplify S5 as an error case of SOCC whose focus is predicted by our CLAN model mistakenly. Here, we present the complete version of S5 in Figure 7a.

B.1.2 PB-FOC

*SEM 2012 Shared Task [Morante and Blanco, 2012] is devoted to negation, specifically, to extract scopes and focuses. Accordingly, *SEM 2012 released two datasets, i.e., CD-SCO for negative scope resolution (i.e., NSR) as task 1, and PB-FOC for negative focus detection (i.e., NFD) as task 2.

In this paper, we select PB-FOC for NFD since it is a widely-accepted focus corpus. This corpus provides focus annotation on top of PropBank. It targets exclusively verbal negations marked with MNEG in PropBank and selects as focus the semantic role containing the most likely focus. The statistics of PB-FOC can be found in Table 3.

B.1.3 CNeSp

CNeSp [Sheng et al., 2020] is a Chinese speculation and negation corpus. In this paper, we use those negation information (scopes and focuses) as the evaluation dataset. CNeSp is comprised of three sub-corpora:

Table 4: Statistics of samples of negation in CNeSp corpus.

(a) CFA

	Scope	Focus		
		Sgl	Oth	All
#Sentences	1260	107	1187	1260
Average Length of Sentences	64.90	69.48	65.02	64.90
#Samples	1614	115	1499	1614
Average Length of Samples	11.15	1.00	4.60	4.34

(b) CPR

	Scope	Focus		
		Sgl	Oth	All
#Sentences	2635	702	2247	2635
Average Length of Sentences	42.87	53.05	44.72	42.87
#Samples	3985	817	3168	3985
Average Length of Samples	7.44	1.00	3.47	2.97

(c) CSL

	Scope	Focus		
		Sgl	Oth	All
#Sentences	132	4	129	132
Average Length of Sentences	70.36	63.00	70.70	70.36
#Samples	161	4	157	161
Average Length of Samples	15.41	1.00	6.50	6.37

Table 5: Statistics of negative scopes in BioScope.

	ABS	CLI	FUL
#Sentences	1597	865	339
Average Length of Sentences	29.28	8.53	30.56
#Samples	1719	870	376
Average Length of Samples	8.60	4.87	8.35

- **Chinese Financial Articles** (**CFA**). The texts of this sub-corpus are from "股市及时雨(timely rain for stock market)" column on Sina.com, mainly involving discussions on stock market trends.
- **Chinese Product Review** (**CPR**). This sub-corpus consists of comments of hotel service from the website Ctrip.com, whose genre is usually free-style, e.g., S8 in Figure 6.
- **Chinese Scientific Literature** (**CSL**). This sub-corpus has articles from Chinese Journal of Computers, an authoritative academic journal in Chinese. Sentences in CSL are usually long and complicated, e.g., S9 in Figure 6. The complete version of S9 can be referred to Figure 7b.

The statistics of the above three sub-corpora is presented in Table 4.

B.2 MORE DETAILS OF EXPERIMENTAL RESULTS

Cross-Domain Detection As mentioned in the last paragraph of § 4.3, to evaluate generalization of CLAN, we get cross-domain performance by treating CNeSp as one Chinese dataset and SOCC&PB-FOC as one English dataset. We clarify that we conduct experiments on Chinese and English datasets in isolation, rather than cross-lingual evaluation.

Performance of BERT We employ NegBERT [Khandelwal and Sawant, 2020], a task-specific model designed for negation detection, and display the performance in Table 6. It can be observed that CLAN and RNet are both better than NegBERT, which can manifest that BERT can not determine the performance of CLAN and RNet, although they use BERT

Table 6: Performance of NegBERT, LLMs, CLAN and other simple models on Negative Scope Resolution (NSR), Negative XScope Resolution (NXSR), Negative Focus Detection (NFD) tasks.

(a) NSR and NXSR

Models	NSR					NXSR
	SOCC	CNeSp				SOCC
		CFA	CPR	CSL	ALL	
DeepSeek-V3	23.45	24.15	21.12	14.91	21.73	3.33
DeepSeek-llm-7B-Chat	25.20	26.84	23.52	15.44	23.84	8.67
Qwen3-8B	27.42	28.32	26.60	16.46	26.77	13.17
BLSTM	51.57	65.69	49.77	36.92	53.51	21.83
NegBERT	63.87	73.22	57.35	47.22	61.49	33.83
RNet	64.71	74.14	58.29	48.34	62.83	31.50
CLAN	68.56	78.51	65.42	47.61	68.58	36.67

(b) NFD

Models	SOCC	PB-FOC	CNeSp			
			CFA	CPR	CSL	ALL
DeepSeek-V3	12.24	15.59	13.03	14.19	5.75	13.64
Sgl	12.57	16.67	14.79	16.70	0.00	16.40
Cue	-	14.05	-	-	-	-
Oth	11.93	16.01	12.90	13.55	5.81	13.11
DeepSeek-llm-7B-Chat	16.97	18.12	15.38	16.36	7.08	15.82
Sgl	17.20	19.79	17.21	16.73	0.00	16.76
Cue	-	16.76	-	-	-	-
Oth	16.85	18.33	15.24	16.27	7.19	15.64
Qwen3-8B	18.84	19.94	16.87	17.44	8.83	17.05
Sgl	20.02	22.92	20.62	18.79	0.00	18.98
Cue	-	17.30	-	-	-	-
Oth	17.46	20.42	16.59	17.07	9.05	16.67
WTG	36.81	67.13	53.88	54.67	24.20	53.58
Sgl	37.63	66.67	30.50	57.81	0.00	54.20
Cue	-	65.41	-	-	-	-
Oth	33.50	67.98	55.66	53.86	25.42	53.46
NegBERT	37.57	66.29	57.19	57.54	26.55	56.56
Sgl	38.93	67.71	36.57	59.28	0.00	56.24
Cue	-	63.78	-	-	-	-
Oth	36.10	67.05	58.75	57.09	27.89	56.63
CLAN	50.38	78.79	65.98	66.36	36.51	65.46
Sgl	54.75	81.25	46.06	70.50	30.00	67.51
Cue	-	77.84	-	-	-	-
Oth	45.61	78.65	67.50	65.29	33.55	64.97

as the initializer for word embeddings. The primary reason for the better performance of CLAN and RNet is that they are integrated with downstream task-specific adapters that can be fine-tuned specially for negation texts.

Performance of LLMs DeepSeek and Qwen are selected as LLM baselines, including:

- DeepSeek-V3, the API-based model;
- DeepSeek-LLM-7B-Chat and Qwen3-8B, the models that can be fine-tuned with LoRA.

and we present the results in Table 6. LLMs performs much worse than other task-specific models. We can consider the error case in Figure 8, which can illustrate the main reasons of the wrong detection of scopes and focuses by LLMs. For scopes, LLMs usually fail to incorporate all the negative texts into the scope, e.g., neglecting negative cues. While for focuses,

S5: *So if you did **not** opt out of Uber's arbitration clause within 30 days of accepting the agreement (and most drivers did not), please CONTACT US so that we can add you to our list of drivers to pursue individual arbitration if we need to do that, so that you can be assured of being covered by the result we obtain. If you accepted an Uber contract within the last 30 days , you can still opt out of the arbitration clause by emailing optout@uber.com and saying you want to opt out of the arbitration clause.*
Prediction: *accepting the agreement*

(a) NFD

S9: 如此则会导致密文数据查询不再具有明文查询特性而造成安全性方面的挑战，且云环境下密文查询有以下需求：面向云端海量数据特征和用户个性化查询需求，有必要提供针对多关键词、多属性请求的高效密文排序查询，充分表达用户查询兴趣，高效返回更能精确满足用户需求且具有重要度区分的所有排序结果或Top-k排序结果，使用户能快速定位与查询最相关密文数据，同时也减小网络传输开销。(*This will result in the fact that cipher data query no longer has features of clear data query that leads to security challenges, and there are following requirements for cipher data query in the cloud environment: for the requirement of query oriented towards features of massive data in the cloud and user personalization, it is necessary to provide efficient ciphertext sorting queries for multi-keyword and multi-attribute requests, fully express user query interests, efficiently return all sorting results or top-k sorting results that can more accurately meet user requirements and have differentiation of importance, enable users to quickly locate and query the most relevant ciphertext data, meanwhile, reduce the cost of network transmission.*)
Prediction: 密文数据查询不再具有明文查询特性而造成安全性(*cipher data query no longer has features of clear data query that leads to security*)

(b) NSR

Figure 7: Error cases with complete versions of sentences predicted by our CLAN model.

S2: *Canada and U.S. are allies in mutual defense , but mutual defense does not mean we get dragged into global wars of power and dominance . his is called the Pierre Trudeau Doctrine. The basis for Canada refusal to join in America war of ideology and hegemony in Vietnam, war of profit, power and religion in Iraq.*
Predicted Scope: *mean we get dragged into global wars*
Predicted Focus: *we get dragged into global wars*

Figure 8: Error cases predicted by DeepSeek.

Table 7: Performance of RNet and CLAN models on Negative Scope Resolution (NSR) in BioScope.

Models	ABS	CLI	FUL	All
RNet	80.51	92.19	63.13	81.72
CLAN	81.63	92.93	63.65	82.64

LLMs do not extract the most dominant negative texts, and tend to detect other negative ones as a part of focus. It can be concluded that although LLMs can extract several negative texts from the sentence, they are still not good at detecting scopes and focuses precisely.

Performance on BioScope corpus BioScope [Vincze et al., 2008, Szarvas et al., 2008] is also a famous dataset annotated with negation in the biomedical domain, and includes three sub-corpora: scientific paper abstracts (ABS), clinical radiology reports (CLI), biological full papers (FUL). Both in-domain and cross-domain evaluations are considered: 1) 10-fold cross validation is employed when evaluating models on ABS; 2) As the test sets, the results on CLI and FUL are predicted by models trained on ABS.

However, BioScope only annotates negative scopes. We still present the performance of our model on BioScope in Table 7, where CLAN is slightly better than RNet, which is analyzed as below: 1) Scopes in CLI are short and easy to detect. Therefore, upper bound of the performance on CLI is nearly reached; 2) Scopes in FUL are long and complex, making NSR quite difficult; 3) ABS is the main target for evaluation, and our CLAN model can outperform RNet, proving the effectiveness of CLAN on common biomedical texts; 4) Scopes in ABS occupy the majority in BioScope, and have different

genres compared with CLI&FUL. Hence, models obtain low performance on FUL in cross-domain evaluation.

C MORE DETAILS OF ACRONYMS

The main acronyms and corresponding complete names are listed as below.

- **Tasks**
  - ND: Negation Detection
  - NSR: Negative Scope Resolution
  - NFD: Negative Focus Detection
  - NXSR: Negative XScope Resolution
- **Datasets**
  - SOCC: Simon Fraser University Opinion and Comments Corpus
  - PB-FOC: PropBank FOCus corpus
  - CNeSp: Chinese Negation and Speculation corpus
  - CFA: Chinese Financial Articles
  - CPR: Chinese Product Review
  - CSL: Chinese Scientific Literature
  - ABS: scientific paper ABStracts in BioScope
  - CLI: CLInical radiology reports in BioScope
  - FUL: biological FULl papers in BioScope
- **Networks**
  - LA: Linear Attentions
  - BLA: Bi-directional Linear Attention
  - SLA: Signed Linear Attention
  - RLA: Recurrent Linear Attention
  - CLAN: Cascade Linear Attention Network
  - CCM: Collapse and Concatenation Module
- **Learning Mechanisms**
  - HRS: Human-like Reading Strategies
  - MAML: Model-Agnostic Meta Learning
  - MTL: Multi-Task Learning
- **Label Scheme**
  - IOM: Input and Output Map
  - BSM: Boundary Shift Map
  - BSSM: Boundary Shift Start Map
  - BSEM: Boundary Shift End Map